

Rethinking machine unlearning for large language models

Year: 2024 | Citations: 190 | Authors: *Sijia Liu, Yuanshun Yao, Jinghan Jia*

Abstract

We explore machine unlearning in the domain of large language models (LLMs), referred to as LLM unlearning. This initiative aims to eliminate undesirable data influence (for example, sensitive or illegal information) and the associated model capabilities, while maintaining the integrity of essential knowledge generation and not affecting causally unrelated information. We envision LLM unlearning becoming a pivotal element in the life-cycle management of LLMs, potentially standing as an essential foundation for developing generative artificial intelligence that is not only safe, secure and trustworthy but also resource-efficient without the need for full retraining. We navigate the unlearning landscape in LLMs from conceptual formulation, methodologies, metrics and applications. In particular, we highlight the often-overlooked aspects of existing LLM unlearning research, for example, unlearning scope, data–model interaction and multifaceted efficacy assessment. We also draw connections between LLM unlearning and related areas such as model editing, influence functions, model explanation, adversarial training and reinforcement learning. Furthermore, we outline an effective assessment framework for LLM unlearning and explore its applications in copyright and privacy safeguards and sociotechnical harm reduction. Machine unlearning techniques remove undesirable data and associated model capabilities while preserving essential knowledge, so that machine learning models can be updated without costly retraining. Liu et al. review recent advances and opportunities in machine unlearning in LLMs, revisiting methodologies and overlooked principles for future improvements and exploring emerging applications in copyright and privacy safeguards and in reducing sociotechnical harms.